Quantitative Techniques for Earth and Atmospheric Sciences

Takamitsu Ito

January 3, 2019

Chapter 1

Data Analysis

The goal of this chapter is to review fundamental aspects of statistics that are commonly applied to Earth and Atmospheric Sciences. Let us start by looking at a specific example, the monthly mean temperature of Atlanta, Georgia, from 1879 to 2015. The data is available from National Weather Service `http://w2.weather.gov/climate/`.[1]

1.1 Distribution

Figure 1.1 shows the plot of annual mean temperature of Atlanta. There are $(2015 - 1879 + 1) = 137$ years of data points for the annual mean temperature, T_i.

[1] For your convenience the data is available at `http://shadow.eas.gatech.edu/~Ito/webdata/EAS2655`

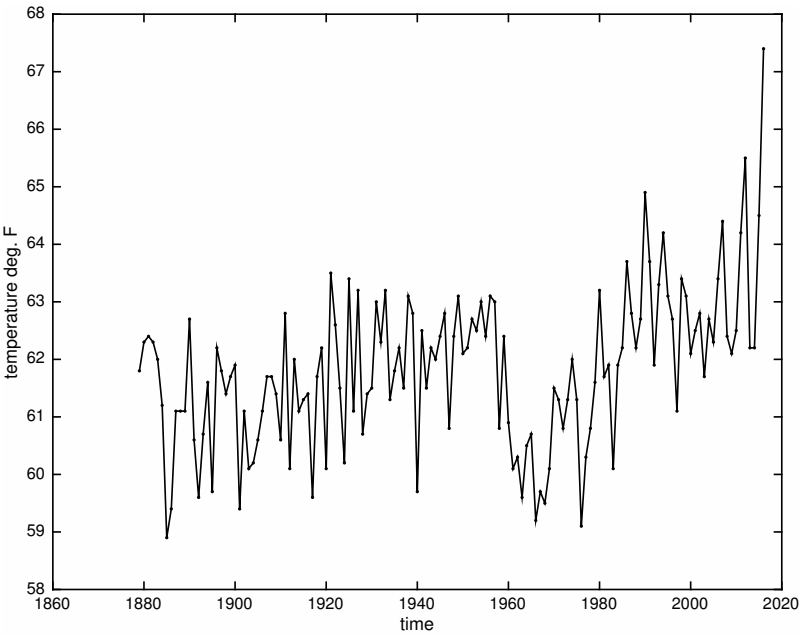

Figure 1.1: Annual mean temperature of Atlanta, Georgia, USA. The plot is made by using MATLAB.

What would be the typical temperature in Atlanta?

The sample mean of a set of T_i can be calculated as

$$\overline{T} = \frac{1}{N} \sum_{i}^{N} T_i, \tag{1.1}$$

where N is the number of samples (that is, 137 for our case). It is summing up the temperatures from all years

and dividing it by the number of years. This is the sample mean or simply average, and is an unbiased estimate of the true mean μ. The mean temperature of Atlanta (1879-2015) is 61.82 degree Fahrenheit. In MATLAB, the mean of variable T can be calculated as follows.

```
>> Tave = mean(T);
```

Median is another way of picking a typical value. It is different from the average. Median is the value that separates the upper half from the lower half of the population. If we line up the 137 temperature values from the coolest to the warmest year, the median temperature would be found as the 69th coolest/warmest temperature, just in the center of the temperature ranking. The median temperature of Atlanta (1879-2015) is 61.9 degree Fahrenheit. In MATLAB, the median of variable T can be calculated using the "median" command.

```
>> Tmedian = median(T);
```

Also, it can be calculated as a 50 percentile of T using the "prctile" command.

```
>> Tmedian = prctile(T,50);
```

In this example, the mean and median are very close to one another. However, that is not always the case. The

mean and median may not take similar values depending on the nature of the variable.

In a skewed distribution, the mean may be more influenced by outliers. Outlier is a rarely observed data point that contains extremely different value from the majority of the distribution. Consider a group of 100 people, where 99 people makes $50,000 and 1 person makes $2 million annually. The average income is $69,500 but that does not really represent the typical income of this group. In fact, there is not even a single person who makes the average income in this group. The median income, however, is $50,000, and is indeed representative of the typical income of this group. Median is known to provide a more robust estimate over different types of distributions.

How much does the temperature vary year to year?

The sample *variance* can be calculated as

$$s^2 = \frac{1}{N-1}\sum_{i}^{N}(T_i - \overline{T})^2. \tag{1.2}$$

An interesting fact is that the above formula of variance use $(N-1)$ instead of N. The detailed derivation can be found in any standard textbook of statistics. The factor $(N-1)$ accounts for the underestimation of the variance due to the uncertainties in the sample mean. In MATLAB,

```
>> Tvar = var(T);
```

calculates the variance of T.

In practice, the variance can be calculated efficiently by taking the difference between the mean of the square and the square of the mean.

$$s^2 = \frac{N}{N-1}(\overline{T_i^2} - \overline{T}^2). \tag{1.3}$$

This translates to a few lines of MATLAB scripts.

```
>> N = length(T);
>> Tave = mean(T);
>> T2ave= mean(T.^2);
>> Tvar = N/(N-1)*(T2ave-Tave);
```

Sample *standard deviation*, s, is calculated as the square root of the sample variance, and is an estimate of the true standard deviation σ. The variance of annual mean temperature of Atlanta (1879-2015) is 1.83 degree Fahrenheit^2, and the standard deviation is 1.35 degree Fahrenheit.

Alternative way of measuring the fluctuation is based on the distribution of the data. Again, if we line up the 137 temperature values from the coolest to the warmest, we can determine a range of temperature for a given percentile. For example, the 25% of data lies below 61.1 degree Fahrenheit and 75% of data lies below 62.7 degree Fahrenheit. Then *Inter-Quartile Range*, IQR, is the

range between the 25% and 75% of the population, that is 1.6 degree Fahrenheit. In MATLAB, we can calculate the IQR as follows.

```
>> up = prctile(T,75);
>> dn = prctile(T,25);
>> Tiqr= up - dn;
```

Alternatively, we can directly calculate IQR using the "iqr" command.

```
>> Tiqr = iqr(T);
```

In the example of the annual mean temperature of Atlanta, the standard deviation and IQR are relatively close to one another. However, the standard deviation and IQR can significantly differ depending on the nature of the distribution. In the previous example of the highly skewed income distribution, the standard deviation of the income is \$195,000 while IQR is \$0. Which number is more representative of the fluctuation? Variance (and standard deviation) takes into account all data points, and as a result it is influenced by the extreme values of the outliers. In contrast, IQR focuses on the 25 to 75 percentile of the distribution only, so it is not influenced by the outliers. Thus IQR is known to provide a more robust estimate of the *typical* range of fluctuations than standard deviation.

Probability distribution

Figure 1.2 is the histogram of the annual mean temperature of Atlanta. This diagram plots how frequently a certain annual mean temperature is observed in this historic dataset. To be more precise, there is a certain range in the temperature (in this case, data is grouped into 1 degree Fahrenheit bins). In MATLAB, you can generate a histogram using the "hist" command.

```
>> hist(T,[55:68]);
```

Here, the histogram of the variable T is generated, using 1 degree bins centered at 55, 56, ... 68 degree F.

For example, there are 45 occurrences of the temperature range between 61.5 to 62.5 degree Fahrenheit. Similarly, there are 35 occurrences between 60.5 to 61.5 degree Fahrenheit, and so on and forth. The histogram visualizes the probability distribution of the temperature data. If we were to randomly select one year from the 137 years of this historic dataset, what is the chance of the temperature to be within 61.5 to 62.5 degree Fahrenheit? The answer is $45/137 = 0.33$. There is one in three chances that the annual mean temperature would end up in the 61.5 to 62.5 degree range.

Theoretically we can construct a continuous probability density function, $f(x)$, such that the probability of

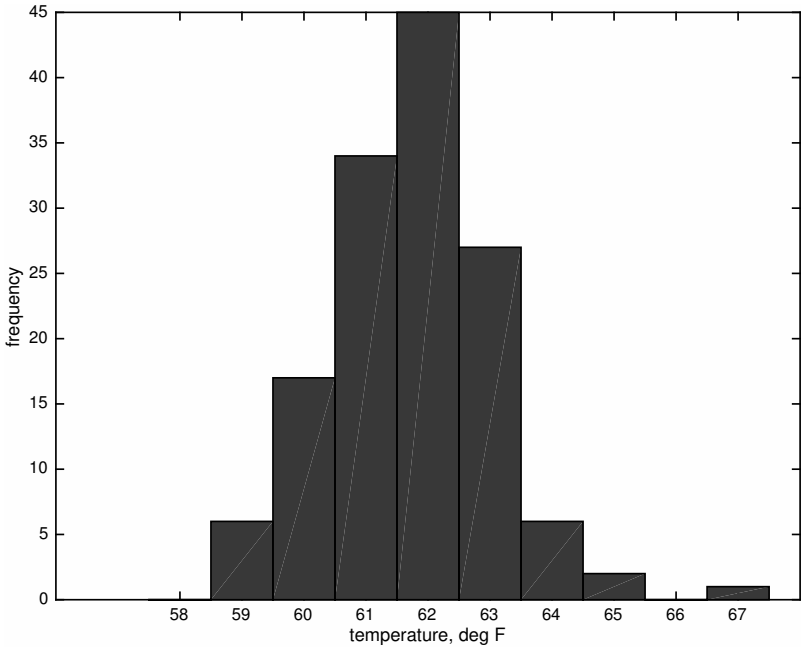

Figure 1.2: Histogram of the annual mean temperature of Atlanta, Georgia, USA. All data is grouped into 1 degree Fahrenheit bin.

the temperature (T) to be bounded by $a < T < b$ is

$$P(a < T < b) = \int_a^b f(x)dx. \tag{1.4}$$

This calculation is exactly analogous to estimating the area of a bar in the histogram (Figure 1.2). The difference is that the area of the bar in the histogram is the actual occurrence, so it is not normalized by the total sample

number. In contrast, the probability density function is normalized, and so the area covered under $f(x)$ is set to 1.

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \tag{1.5}$$

Cumulative distribution function is the integral of the probability density function, and it calculates the probability that the temperature would be below T.

$$P(x < T) = F(T) = \int_{-\infty}^{T} f(x)dx. \tag{1.6}$$

In relation the calculation of the probability in Eq. 1.12, the probability of observing temperature bounded by a and b would be

$$P(a < T < b) = F(b) - F(a). \tag{1.7}$$

Gaussian, normal distribution

In many cases, Gaussian (normal distribution) provides a good estimate of the probability density for environmental variables. It is a bell-shaped curve centered at zero with a standard deviation of one.

$$f(x) = \frac{1}{\sqrt{2\pi}} exp(-\frac{x^2}{2}) \tag{1.8}$$

The Gaussian can also be written for a variable with non-zero mean and non-unity standard deviation.

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{1.9}$$

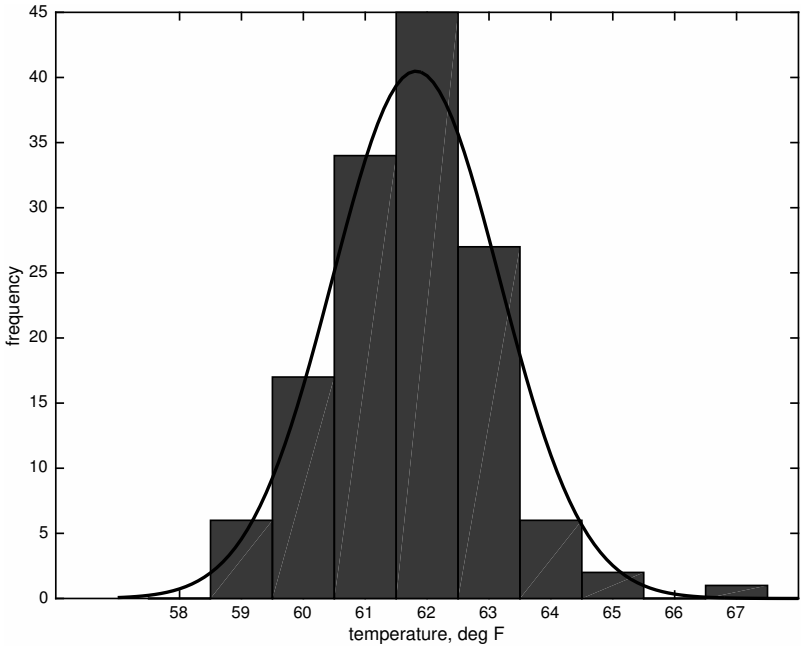

Figure 1.3: Histogram of the annual mean temperature of Atlanta, Georgia, USA, overlain with the scaled Gaussian distribution

The factor multiplying the above exponentials ensures that the integral of f is 1 so it satisfies Eq. 1.5.

Let's compare the distribution of the annual mean temperature of Atlanta to the Gaussian distribution. We use the formula of Eq. 1.9 with the mean and the standard deviation calculated earlier. Figure 1.3 compares the histogram of the temperature and the Eq. 1.9 multiplied by N (=137) so its area is appropriately scaled.

The Gaussian function (solid line in Figure 1.3) captures quite well the overall shape of the histogram. Then we can say that the annual mean temperature of Atlanta approximately follows *normal distribution*.

Standardization

Any variable can be standardized to have a zero mean and one standard deviation.

$$z = \frac{x - \mu}{\sigma} \tag{1.10}$$

z is the standardized variable, and then Eq. 1.8 can be compared to the distribution of z. The standardization is a useful technique because it removes the magnitudes and units from the variable, and it allows us to look at its fluctuations only. There are some notable properties of Gaussian distribution.

$$P(-1 < z < +1) = \int_{-1}^{+1} f(x)dx = 0.68 \tag{1.11}$$

If a variable is normally distributed, approximately two thirds of the population remains within ± 1 standard deviation about the mean. So there is about one in three chance that a random selection can result in outside of ± 1 standard deviation about the mean.

$$P(-2 < z < +2) = \int_{-2}^{+2} f(x)dx = 0.95 \qquad (1.12)$$

If we consider a wider envelope of ± 2 standard deviation about the mean, approximately 95% of population is contained within this range. That is one in twenty chance that a randomly selected variable can be outside of the ± 2 standard deviation.

Box-Whiskar plot

In addition to the histogram (Fig 1.3), box-whiskar plot is a useful way of visualizing the statistical distribution. It visualizes five statistics that characterizes the data. The box-whiskar plot of the Atlanta temperature is shown in Fig 1.4. First, it shows a box, centered at the median, and the two sides are set to the 25 and 75 percentiles. In another words, the width of the box is the IQR. There are two lines (whiskars) extending from the sides of the box, indicating the minimum and maximum values.

The box-whiskar diagram (Fig 1.4) was generated by the "boxplot" command in MATLAB.

```
>> boxplot(T);
>> ylabel('Temperature, deg F');
```

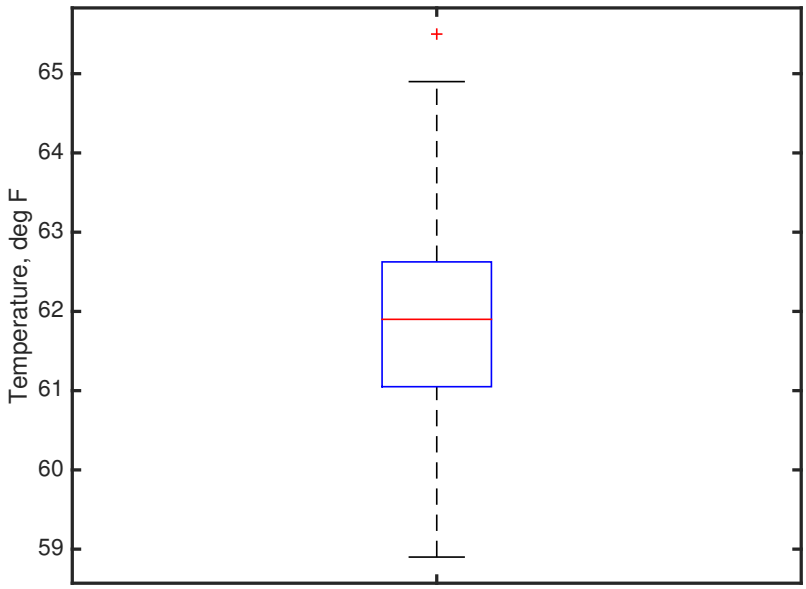

Figure 1.4: Box-Whiskar plot of the annual mean temperature of Atlanta, Georgia, USA. The red cross indicates outlier(s). The box is centered at median value, and its vertical extent is the IQR bounded by 25 and 75 percentile values. Whiskar indicates the maximum and minimum values.

Exercises

1. Download the monthly mean temperature of Atlanta from the course website, `http://shadow.eas.gatech.edu/~Ito/webdata/EAS2655`.

2. Calculate and display the mean, median, standard deviation and IQR of the July temperature.

3. Reproduce Figure 1.1 and 1.3 for the month of July.

4. Climatology refers to the long-term mean values. Display the monthly statistics of Atlanta's temperature by plotting Box-Whiskar diagram for each of the 12 months and plot them together in a single panel using month as the x-axis.

5. **HW1** Publish the MATLAB script that performs all of the activity above, and submit it as a report in the PDF format.